BERJ British Educational Research Journal

BERA BRITISH EDUCATIONAL RESEARCH ASSOCIATION

British Educational Research Journal
Vol. 45, No. 3, June 2019, pp. 518–537

DOI: 10.1002/berj.3511

The importance of adjusting for pupil background in school value-added models: A study of Progress 8 and school accountability in England

George Leckie* and Harvey Goldstein
Centre for Multilevel Modelling and School of Education, University of Bristol, UK

In the UK, USA and elsewhere, school accountability systems increasingly compare schools using value-added measures of school performance derived from pupil scores in high-stakes standardised tests. Rather than naïvely comparing school average scores, which largely reflect school intake differences in prior attainment, these measures attempt to compare the average progress or improvement pupils make during a year or phase of schooling. Schools, however, also differ in terms of their pupil demographic and socioeconomic characteristics and these factors also predict why some schools subsequently score higher than others. Many therefore argue that value-added measures unadjusted for pupil background are biased in favour of schools with more 'educationally advantaged' intakes. But others worry that adjusting for pupil background entrenches socioeconomic inequities and excuses low-performing schools. In this article we explore these theoretical arguments and their practical importance in the context of the 'Progress 8' secondary school accountability system in England, which has chosen to ignore pupil background. We reveal how the reported low or high performance of many schools changes dramatically once adjustments are made for pupil background, and these changes also affect the reported differential performances of regions and of different school types. We conclude that accountability systems which choose to ignore pupil background are likely to reward and punish the wrong schools and this will likely have detrimental effects on pupil learning. These findings, especially when coupled with more general concerns surrounding high-stakes testing and school value-added models, raise serious doubts about their use in school accountability systems.

Keywords: Attainment 8; national pupil database; Progress 8; school accountability; school league tables; school performance measures; value-added

Introduction

In the UK, USA and elsewhere, education systems increasingly hold schools to account using school performance measures derived from pupil scores in high-stakes standardised tests and examinations (OECD, 2008; Koretz, 2017; NFER, 2018). Schools are held accountable for the progress or improvement shown by their pupils over a year or phase of schooling. The implicit assumption is that variation in school average progress is a valid indicator of the value that schools add to pupil learning. In other words, the education effectiveness or quality of schools.

*Corresponding author. Centre for Multilevel Modelling, School of Education, University of Bristol, 35 Berkeley Square, Bristol BS8 1JA, UK. E-mail: g.leckie@bristol.ac.uk

England has been at the forefront of this move to test-based school accountability (West, 2010). Successive governments over the last 25 years have introduced new and supposedly improved school performance measures that purport to measure what is happening in schools (Kelly & Downey, 2010a; Leckie & Goldstein, 2017). These measures are also used to promote parental choice via their high-profile publication in 'school league tables' (Leckie & Goldstein, 2009). They are also used by schools for self-evaluation, improvement, tracking and target-setting purposes, with schools increasingly buying in data analysis support from commercial organisations to assist them in these endeavours (Selfridge, 2018, p. 40). The measures also inform national debates around regional inequalities, the performance of different school types and performance gaps across socioeconomic, ethnic and other pupil groups.

In 2016, the government introduced a new secondary school accountability system for all mainstream state-funded schools in England (DfE, 2018c). Attainment 8—essentially a total score across eight subjects—was introduced as the new headline measure of pupil performance at the end of secondary schooling General Certificate of Secondary Education (GCSE; age 15/16). Progress 8—a type of value-added approach—was introduced as the new headline measure of progress or the improvement that pupils make between the end of primary schooling key stage 2 tests (KS2; age 10/11) and the GCSE examinations. Each pupil's score is calculated as their Attainment 8 score minus the average Attainment 8 score of all pupils nationally with the same KS2 prior attainment (KS2 scores are categorised into 34 groups for this purpose). A school's Progress 8 score is simply the average of their pupils' scores and is presented with a 95% confidence interval to communicate its statistical uncertainty. The government argues that Progress 8 leads to fairer and more meaningful comparisons for school accountability purposes than Attainment 8 as it adjusts for school intake differences in KS2 prior attainment. Specifically, schools are labelled 'underperforming' if their Progress 8 scores fall below a minimum standard for progress, referred to as a 'floor standard'. Such schools come under increased scrutiny and intervention from Ofsted, the national school inspectorate, and by regional schools commissioners and local authorities in their roles supporting schools. In contrast, schools with the highest Progress 8 scores are exempt from routine inspections by Ofsted in the following calendar year, a highly desirable outcome for any school.

The design of all school value-added measures and accountability systems is based on subjective modelling decisions and assumptions and given the high stakes nearly always involved, these choices must be independently and robustly evaluated. In this article, we explore a particularly divisive decision relevant to not just Progress 8, but all school value-added measures and accountability systems, namely whether to adjust for school intake differences in pupil demographic and socioeconomic background characteristics since these factors also predict why some schools subsequently score higher than others. For Progress 8, the government has chosen to ignore pupil background. We assess the practical importance of this decision for school accountability in England. We examine in detail the extent to which schools' Progress 8 scores, ranks and classifications as successful and failing schools change when we account for pupil background. We highlight those schools which would

benefit and lose most from any change to Progress 8. We then draw attention to further statistical issues with Progress 8 which demand further research as well as our reservations more generally with regard to test-based school accountability.

To adjust or not to adjust?

Progress 8 adjusts pupils' Attainment 8 scores for their KS2 prior attainment scores but does not adjust for other pupil characteristics which also differ across schools. While prior attainment is nearly always the most important predictor of current attainment in school value-added models, many national and international studies have long shown the secondary importance of pupil demographic and socioeconomic characteristics as additional predictors (Teddlie & Reynolds, 2000; Reynolds *et al.*, 2014). It follows that, in the absence of any adjustments, different pupil groups will typically show different average progress during schooling. Thus, in England, girls typically make more progress during secondary schooling than boys, many ethnic minority groups make more progress than White British pupils, pupils with no special education needs make more progress than those with such needs, and rich pupils make more progress than poor pupils (EPI, 2017). It follows that schools with more 'educationally advantaged' intakes in England would in general be expected to show higher average pupil Progress 8 scores than schools with less educationally advantaged intakes.

Some studies have additionally shown that school average pupil prior attainment and, to a lesser extent, various school average background characteristics also predict subsequent pupil attainment even after adjusting for the pupil versions of these variables (Timmermans & Thomas, 2015). However, we do not consider the possibility of 'school compositional effects' further in this article as, while potentially important, there are several unresolved statistical issues connected to attempts to adjust for such effects related to confounding (systematic sorting of more advantaged pupils into more effective schools) and measurement error (prior attainment an unreliable measure of true attainment) (Castellano *et al.*, 2014; Perry, 2018).

The argument for adjusting: To make fair and meaningful comparisons

Many academics and educationalists argue that failing to adjust for pupil background is fundamentally unfair as it punishes some schools merely for teaching educationally disadvantaged intakes and rewards other schools merely for teaching educationally advantaged intakes (Raudenbush & Willms, 1995; Goldstein, 1997; Teddlie & Reynolds, 2000; OECD, 2008; Reynolds *et al.*, 2014; BBC, 2018; TES, 2018). The true effectiveness of many schools in disadvantaged areas will go undetected, as will the lack of effectiveness or 'coasting' performance of many schools in advantaged areas. School value-added measures such as Progress 8, which ignore pupil background, are therefore likely to punish and reward the wrong schools and to hold up the wrong schools as examples of success that other schools should learn from. Furthermore, punishing schools for teaching disadvantaged pupils is likely to incentivise schools to avoid admitting particular pupil groups (e.g. children with special educational needs), and where they are admitted, to find ways to exclude them from the examinations and therefore the value-added calculations. Indeed, in

England, there has been a rise in pupil exclusions over the last 2 years, which in part has been attributed to schools gaming the accountability system in these ways (DfE, 2018a). A related concern is that unadjusted school value-added measures require disadvantaged pupils in each school to make as much progress as their advantaged peers. However, given the differential performance of many pupil groups, this is simply an unrealistic target, at least in the short run, and so is likely to leave many disadvantaged pupils and their schools feeling as if they have failed. This may dissuade good teachers from working in challenging schools and may induce teachers in those schools to leave. Proponents of all these arguments therefore argue that school value-added measures must adjust not just for prior attainment but additionally for pupil socioeconomic status and other pupil characteristics that predict subsequent attainment.

The argument against adjusting: It lowers expectations of disadvantaged groups

Others argue against adjusting school value-added measures for pupil background, worrying that such adjustments entrench socioeconomic inequities and excuse low-performing schools. Indeed, this argument was made by the government when they withdrew the previous administration's 'contextual value-added' measure, which did adjust for pupil background (DfE, 2010; Leckie & Goldstein, 2017). In terms of Progress 8, the UK government continues to argue that society should expect disadvantaged pupils with the same prior attainment as their more advantaged peers to continue to perform at the same academic level at GCSE, not fall behind (Burgess & Thomson, 2013). There is, however, a lack of any theoretical justification for such an assertion. Moreover, it seems inconsistent to acknowledge the empirical fact that pupils from disadvantaged backgrounds are already behind when they start their secondary schooling, but to refuse to accept the empirical fact that this 'deficit' is not fully removed by adjusting for their lower prior attainment.

The government goes on to argue that adjusting for the lower progress of disadvantaged pupil groups entrenches low aspirations for these pupils (DfE, 2010). However, if one accepts this argument then one must also accept that adjusting for prior attainment entrenches low aspirations for low-prior-attaining pupils. Thus, using this argument to ignore pupil background but to adjust for pupil prior attainment appears inconsistent (Perry, 2016).

One practice that may entrench low aspirations for particular pupil groups is the widespread practice of target setting in schools, since here empirical relationships between attainment and pupil background characteristics in previous school cohorts is used to predict the future performance of current pupils (Castellano & Ho, 2013; Selfridge, 2018). When used in a deterministic fashion, this may propagate past inequities onto future generations (Leckie & Goldstein, 2017). However, others would argue that when used properly, target setting can be used to show schools that many pupils outperform the average performance of their groups and this can then encourage schools to work harder for these groups. Empirical studies are clearly needed to study the effect of target setting on pupil attainment and progress, and more generally on the impact of commercial organisations providing this type of data analysis support to schools.

Data

We focus on the 3,098 schools whose Progress 8 scores were published in the government's 2016 secondary school performance tables: essentially all state-maintained secondary schools in England. We use the government's National Pupil Database to recreate the underlying pupil-level Attainment 8 and KS2 score dataset from which school Progress 8 scores are derived. We additionally merge in a range of standard pupil background and school characteristics. Pupil characteristics include: age, gender, ethnicity, language (whether they speak English as an additional language), SEN (special educational needs status), FSM (eligible for free school meals at some time in the preceding 6 years: an indicator of poverty) and deprivation (deprivation of the pupil's residential neighbourhood as proxied by the IDACI decile of their home postcode). School characteristics include: region, type, admissions policy, age range, gender, religious denomination and deprivation (deprivation of the school neighbourhood). The final analysis sample consists of 502,851 pupils in 3,098 schools located in 151 local authorities across the nine regions of England.

Table 1 presents pupil- and school-level summary statistics for Progress 8. See the Supporting Information (Figures S1–S3) for plots and equivalent summary statistics and plots for Attainment 8 and KS2 prior attainment. A pupil Progress 8 score of 1.00 corresponds to that pupil scoring 1.00 grade higher per GCSE subject than pupils nationally with the same KS2 prior attainment. Pupil Progress 8 scores are approximately normally distributed with a national mean and SD of 0.00 and 1.06. The mean and SD of school average Progress 8 scores are −0.03 and 0.40, and its distribution is also approximately normal.

Table 2 presents school Progress 8 'bandings'. Essentially, the government assigns each school to one of five bands as a function of the magnitude and statistical significance of their Progress 8 score (DfE, 2018b; see Table 2 for the exact definition of each banding). We see that 303 schools nationally (9.8% of all schools) are assigned to the 'well below average' banding and therefore do not meet the government's minimum standard of progress (defined as the threshold between this banding and the 'below average' banding). In contrast, 193 schools nationally (6.2%) are assigned to the 'well above average' banding.

Table 1. Pupil- and school-level summary statistics for Progress 8 and Adjusted Progress 8

Description	Mean	SD	Min	10th	25th	50th	75th	90th	Max
Pupils (N = 502,851)									
Progress 8	0.00	1.06	−7.39	−1.25	−0.52	0.11	0.69	1.18	5.57
Adjusted Progress 8	0.00	0.99	−7.34	−1.17	−0.51	0.09	0.63	1.11	5.44
Schools (N = 3,098)									
Progress 8	−0.03	0.40	−3.54	−0.50	−0.23	0.00	0.24	0.43	1.37
Adjusted Progress 8	−0.01	0.35	−3.19	−0.40	−0.20	0.01	0.20	0.38	1.30

Note: 10th, 25th, 50th, 75th and 90th denote percentiles of the relevant score distributions.

Table 2. School Progress 8 and school Adjusted Progress 8 bandings

Banding	Definition		Number and % of schools	
	Score	Significant	Progress 8	Adjusted Progress 8
5 = Well above average	≥5	Yes	193 (6.2%)	148 (4.8%)
4 = Above average	>0 & <0.5	Yes	764 (24.7%)	783 (25.3%)
3 = Average		No	1213 (39.2%)	1278 (41.3%)
2 = Below average	≥−0.5 & <0	Yes	625 (20.2%)	693 (22.4%)
1 = Well below average	<−0.5	Yes	303 (9.8%)	196 (6.3%)

Note: Definitions reproduced from DfE (2018b).
Significant = whether the score is significantly different from 0.
Number of schools = 3,098.

The relationship between Progress 8 and pupil background characteristics

In this section, we reveal the very different average pupil progress made by different pupil groups according to Progress 8. Figure 1 (left-hand panel) presents average pupil Progress 8 by pupil age, gender, ethnicity, language, SEN, FSM and deprivation. The categories within each pupil characteristic are sorted by average pupil Progress 8 and for each pupil characteristic the overall variation across the categories is statistically significant (one-way ANOVA tests robust to school-level clustering all show $p < 0.001$). These statistics are preliminary descriptive statistics which analyse each pupil characteristic separately. Later, we will model pupil progress jointly in terms of all seven characteristics. See Supporting Information for the number of pupils in each category of each pupil characteristic (Table S2) and for corresponding plots for Attainment 8 and KS2 prior attainment (Figure S4).

August-born pupils make 0.19 grades more progress per subject than their September-born peers. Given that the SD in pupil Progress 8 is 1.06, this difference is substantial, almost one-fifth of 1.00 SD. More generally, younger pupils within the academic year make more progress than older pupils. However, younger pupils score lower than older pupils at the end of primary schooling and they still do so at the end of secondary schooling, despite their higher progress (Supporting Information: Figure S4). Thus, the higher progress shown among younger pupils reflects their attainment approaching, but not reaching, the higher attainment of their older peers during secondary schooling. These patterns agree with Crawford *et al.* (2007) and others who have done work on month-of-birth effects in England.

Girls make 0.26 grades more progress per subject than boys. However, girls already score higher than boys at the end of primary schooling (Supporting Information: Figure S4) and so the end-of-primary-school gender attainment gap widens over secondary schooling. Potential explanations are discussed in detail by Sammons (1995), among others.

There is substantial variation in Progress 8 by ethnic group. Chinese pupils (0.3% of all pupils) score, on average, 0.70 grades higher per subject than expected given their prior attainment; Indian pupils (2.5%), 0.49 grades higher; Black African pupils (2.9%), 0.37 grades higher; and Bangladeshi pupils (1.5%), 0.35 grades higher. In contrast, White British pupils (76%), on average, score 0.08 grades lower than

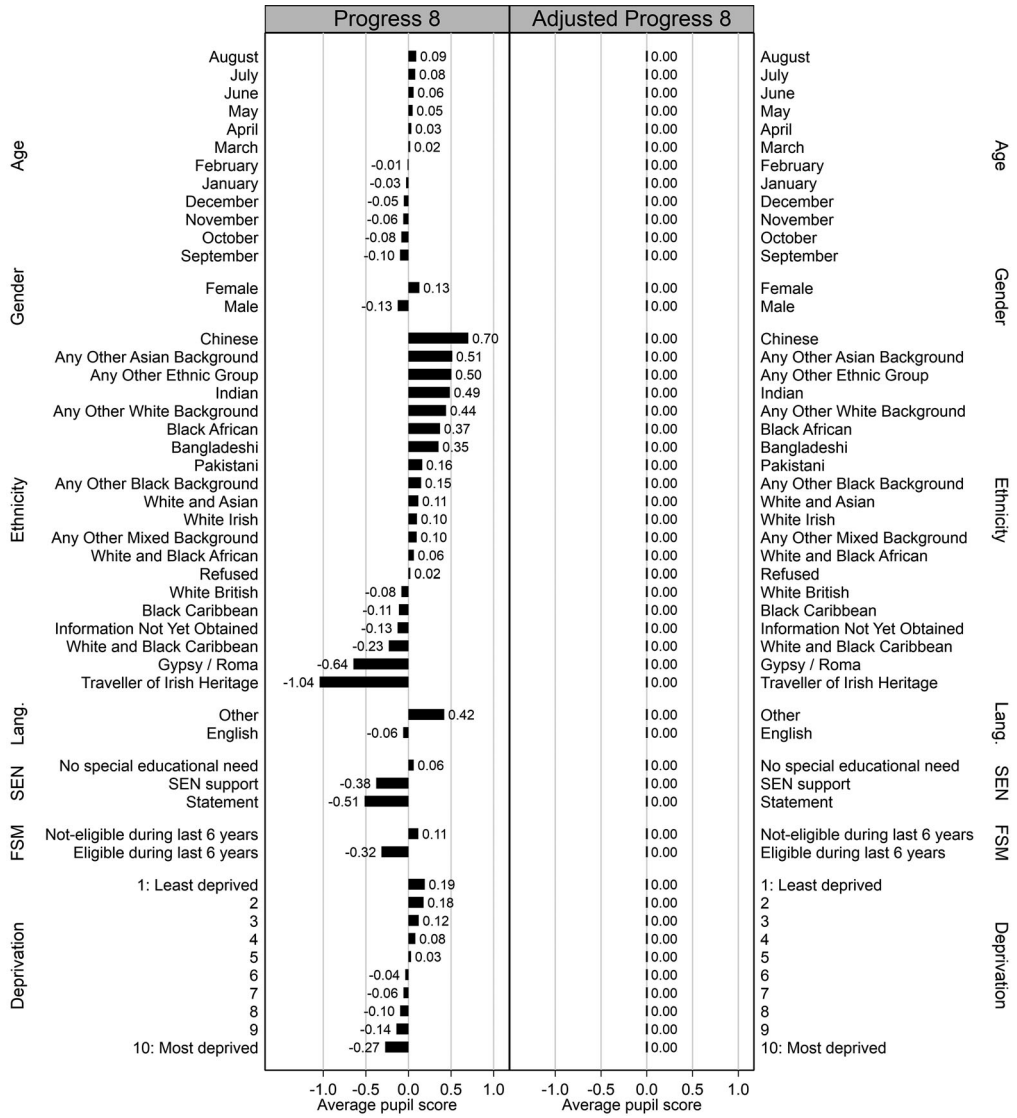

Figure 1. Average pupil Progress 8 and Adjusted Progress 8 scores by pupil characteristics
Note: By definition, there is no variation in average Adjusted Progress 8 by pupil characteristic. The numbers of pupils by pupil characteristic are given in Table S2.

expected. Black Caribbean pupils (1.3%) do worse still, scoring 0.11 grades lower than expected. However, Gypsy/Roma pupils (0.1%) and Travellers of Irish Heritage (0.02%) show the lowest progress, scoring 0.64 and 1.04 grades lower. These progress gaps in England are long-standing and their causes are complex and intertwined with the differing socioeconomic status and other characteristics of these groups (Wilson *et al.*, 2011; Strand, 2014).

Pupils speaking English as an additional language (13% of all pupils) make 0.48 grades more progress per subject than pupils who speak English as their first

language. Essentially, this pupil group catches up and by the end of secondary schooling overtakes their peers who speak English as a first language (Supporting Information: Figure S4). Strand *et al.* (2015) describe in detail the relationships between pupil attainment, progress and language status in England.

Pupils with SEN support (11% of all pupils), especially those with statements (2%), make considerably less progress than pupils with no special education needs. These two pupil groups already score lower at the end of primary schooling and so these attainment gaps widen during secondary schooling (Supporting Information: Figure S4).

Pupils eligible for FSM (27% of all pupils) make 0.43 grades less progress per subject than pupils who are not eligible for FSM. Ilie *et al.* (2017) provide a recent discussion of FSM differences in progress, including the strengths and weaknesses of using FSM as a proxy for socioeconomic disadvantage.

Pupils residing in disadvantaged neighbourhoods also make less progress than those in more prosperous neighbourhoods. For example, pupils living in the most affluent 10% of neighbourhoods score, on average, 0.19 grades higher per subject than predicted by their prior attainment, while pupils living in the poorest 10% of neighbourhoods score 0.27 grades lower per subject than predicted. This social gradient is already present at the end of primary schooling and so widens over secondary schooling (Supporting Information: Figure S4).

Modifying Progress 8 to adjust for pupil background characteristics

In this section, we modify Progress 8 to adjust for the seven pupil background characteristics described above: age, gender, ethnicity, language status, SEN, FSM and deprivation. We refer to this measure as 'Adjusted Progress 8'.

Recall that each pupil's Progress 8 score is calculated as their actual Attainment 8 score minus the average Attainment 8 score across all pupils nationally with the same KS2 prior attainment, where KS2 prior attainment is categorised into 34 bands for this purpose. The calculation of pupil and school Progress 8 scores can therefore be viewed as an application of linear regression. Essentially, pupil Progress 8 scores are calculated as the residuals from a linear regression of pupil Attainment 8 on 34 dummy variables, one for each KS2 band. School Progress 8 scores are then calculated as school averages of these residuals. This reformulation reveals the government's approach to be at odds with the considerable methodological and applied research literature on measuring school effects which favours a multilevel modelling approach, a point we return to in the discussion later (Aitkin & Longford, 1986; Raudenbush & Willms, 1995; Goldstein, 1997, 2011; Teddlie & Reynolds, 2000; OECD, 2008; Reynolds *et al.*, 2014).

We explore the importance of adjusting for pupil background on Progress 8 as simply as possible by entering these seven pupil characteristics into the Progress 8 linear regression model. Thus, we retain all other features of the government's methodology. We do not include interaction terms as the use of the 34 dummy variables for prior attainment means that interactions between prior attainment and the pupil characteristics would result in a very large number of parameters, many of which would be poorly estimated. Given the importance of accounting for such interactions

(Goldstein, 1997), this is a clear limitation of the Progress 8 methodology (it would seem preferable to enter prior attainment as a low-order polynomial). Figure 1 (right-hand panel) confirms that the Adjusted Progress 8 model fully adjusts for the seven pupil characteristics: the average pupil progress for every pupil group is now 0.00.

The full results for the Progress 8 and Adjusted Progress 8 models can be found in the Supporting Information (Table S4). Here we summarise the overall fit of these two models to the data. The Progress 8 model results in 34 regression coefficients. The adjusted R-squared is 0.570 and so pupils' KS2 scores predict 57% of the variation in their Attainment 8 scores. In contrast, the Adjusted Progress 8 model results in 78 regression coefficients and an increased adjusted R-squared of 0.624. The standard deviation of pupils' progress scores reduces by 6.6% while the correlation between the pupil Adjusted Progress 8 scores and pupil Progress 8 is 0.895. These statistics suggest that while prior attainment is clearly the most important predictor of Attainment 8, the seven pupil characteristics nonetheless improve these predictions.

Comparing Progress 8 and Adjusted Progress 8 scores, ranks and bandings

In this section we reveal the practical importance of adjusting for pupil background by comparing Progress 8 and Adjusted Progress 8 scores, ranks and classifications.

Reconsider Table 1. Focusing on the school-level statistics, the means of both variables are effectively zero, but the standard deviation (SD) of Adjusted Progress 8 is lower than that of Progress 8 (0.35 vs. 0.40). Thus, school Adjusted Progress 8 scores are in general smaller in absolute value than school Progress 8 scores. The intuition is that Progress 8 overstates the effects schools have on their pupils: part of the measured effects simply reflects school intake differences in pupils' backgrounds.

Figure 2 presents scatterplots of school Attainment 8, Progress 8 and Adjusted Progress 8 scores (first row) and ranks (second row). The Progress 8 against Attainment 8 scatterplots (first column) suggest schools with the best Attainment 8 results tend, but are no means guaranteed, to be the schools where pupils make the most progress (Pearson correlation: $r = 0.75$; Spearman rank correlation: $r_s = 0.77$). The small cluster of schools distinct from the rest (top plot) are grammar schools whose unusual performance we shall return to later. The Adjusted Progress 8 against Attainment 8 scatterplots (second column) show a somewhat weaker relationship ($r = 0.61$; $r_s = 0.62$), illustrating again that part of what is measured by Progress 8 is school variation in pupil background. The Adjusted Progress 8 against Progress 8 scatterplots (third column) show the strongest associations ($r = 1.91$; $r_s = 0.89$). However, even here, school performance differs greatly depending on which progress measure schools are judged by. This is shown by the substantial number of schools located away from the 45-degree line in the bottom plot. Indeed, changing from Progress 8 to Adjusted Progress 8 would lead 574 schools (19% of all schools in the country) to move up or down the national league table by 500 or more ranks, with 110 schools (4%) moving over 1,000 ranks. Bearing in mind that there are only around 3,000 secondary schools nationally, these changes are very large indeed.

Table 3 presents a cross-tabulation of school Progress 8 bandings (rows) and Adjusted Progress 8 bandings (columns). The row percentages present the percentage of schools within each Progress 8 banding that are assigned to each Adjusted

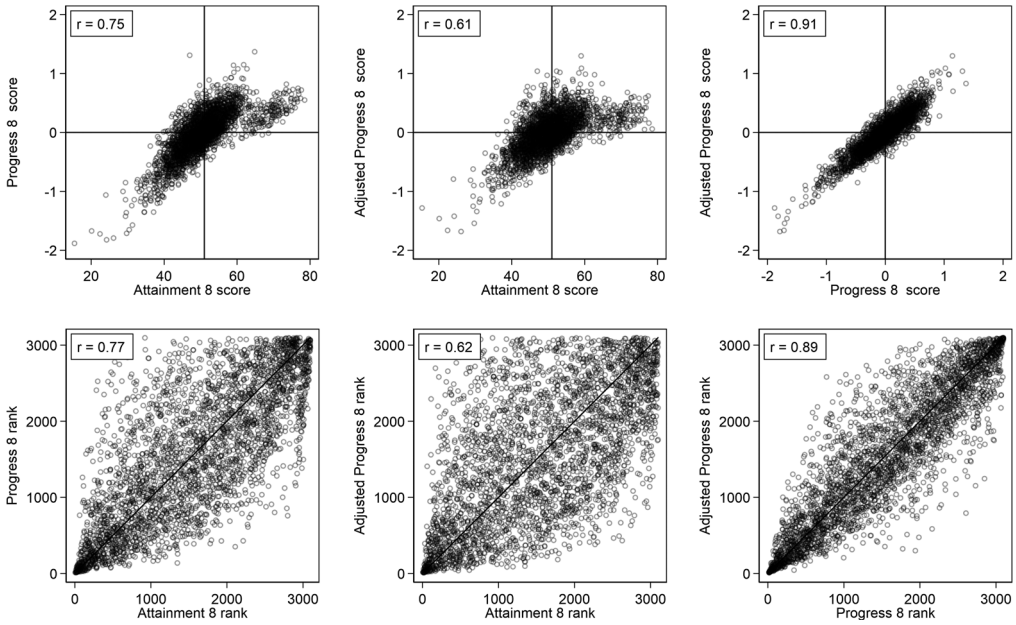

Figure 2. Scatterplots of school average Attainment 8, Progress 8 and Adjusted Progress 8 scores (first row) and ranks (second row) with Pearson and Spearman rank correlations
Note: The horizontal and vertical lines in the first row of plots denote the mean values of the relevant variables.

Progress 8 banding. The table shows that moving from Progress 8 to Adjusted Progress 8 would lead 988 schools (32% of all schools) to change bandings. Importantly, the number of schools assigned to the 'well below average' banding and therefore judged to be performing below the government's floor standard would drop from 303 schools (9.8% of all schools) to 196 schools (6.3% of all schools), a decrease of 107 schools, or just over a third. At the other extreme, the number of schools assigned to the 'well above average' banding would decrease from 193 schools (6.2% of all schools) to 148 schools (4.8% of all schools), a decrease of 45 schools, or almost a quarter.

The decrease in the number of schools appearing in these two most extreme bandings is consistent with the lower SD reported for school Adjusted Progress 8 scores compared to school Progress 8 scores (0.35 vs. 0.40; Table 1). The intuition is that by setting more realistic expected Attainment 8 scores for pupils, fewer pupils would be deemed to make irregular progress and so fewer schools would be judged to be substantially under- or overperforming and therefore appearing in the two most extreme bandings. However, this is not to imply that no schools would move into the two most extreme bandings under Adjusted Progress 8. Indeed, 16 schools judged 'below average' under Progress 8 would be judged 'well below average' under Adjusted Progress 8 and therefore now in line for Ofsted intervention. The intuition here is that the previously acceptable average pupil progress seen in these schools is no longer acceptable once we learn that these schools disproportionately teach educationally advantaged pupils.

Table 3. Cross-tabulation of school Progress 8 bandings by school Adjusted Progress 8 bandings

	Adjusted Progress 8 banding					
Progress 8 banding	Well below	Below	Average	Above	Well above	Total
Well above	0	0	5	101	87	193
	0.0%	0.0%	2.6%	52.3%	45.1%	100%
Above	0	3	195	511	55	764
	0.0%	0.4%	25.5%	66.9%	7.2%	100%
Average	0	141	898	168	6	1,213
	0.0%	11.6%	74.0%	13.9%	0.5%	100%
Below	16	434	172	3	0	625
	2.6%	69.4%	27.5%	0.5%	0.0%	100%
Well below	180	115	8	0	0	303
	59.4%	38.0%	2.6%	0.0%	0.0%	100%
Total	196	963	1,278	783	148	3,098
	6.3%	22.4%	41.3%	25.3%	4.8%	100%

Note: Definitions of bandings are given in Table 2.

Comparing Progress 8 and Adjusted Progress 8 scores by school characteristics

In this section, we describe which types of schools would, on average, benefit or lose from any move to adjust Progress 8 for pupil background. We do this by comparing pupil average Progress 8 and Adjusted Progress 8 scores by school region, type, admissions policy, age range, gender, religious denomination and deprivation.

The left- and right-hand panels of Figure 3 present pupil average Progress 8 and Adjusted Progress 8 scores by each school characteristic in turn. To facilitate comparison, the categories within each school characteristic, for both measures, are sorted by average pupil Progress 8 scores. In every case, the variation across the categories of each school characteristic is statistically significant (one-way ANOVA tests robust to school-level clustering all show $p < 0.001$). As with Figure 1, these are simple descriptive statistics which analyse each characteristic separately. See Supporting Information for the number of pupils and schools by each school characteristic (Table S3) and for corresponding plots for Attainment 8 and KS2 prior attainment (Figure S5).

According to Progress 8 (left-hand panel), pupils in London schools (431 schools; 14% of all schools) make, on average, the most progress, scoring 0.19 grades higher per subject than pupils nationally with the same prior attainment. However, under Adjusted Progress 8 (right-hand panel) this 'London effect' halves to just 0.09 grades per subject. Further analysis suggests that while London schools are somewhat disadvantaged by teaching relatively poor intakes (they have relatively high rates of FSM pupils and pupils in deprived neighbourhoods), they are to a much greater extent advantaged by teaching particular ethnic groups who nationally tend to make high progress (in particular, Black Africans, Any Other Ethnic Group, Any Other White Background, Bangladeshi and Indian). They also teach high proportions of pupils who speak English as an additional language, another high-

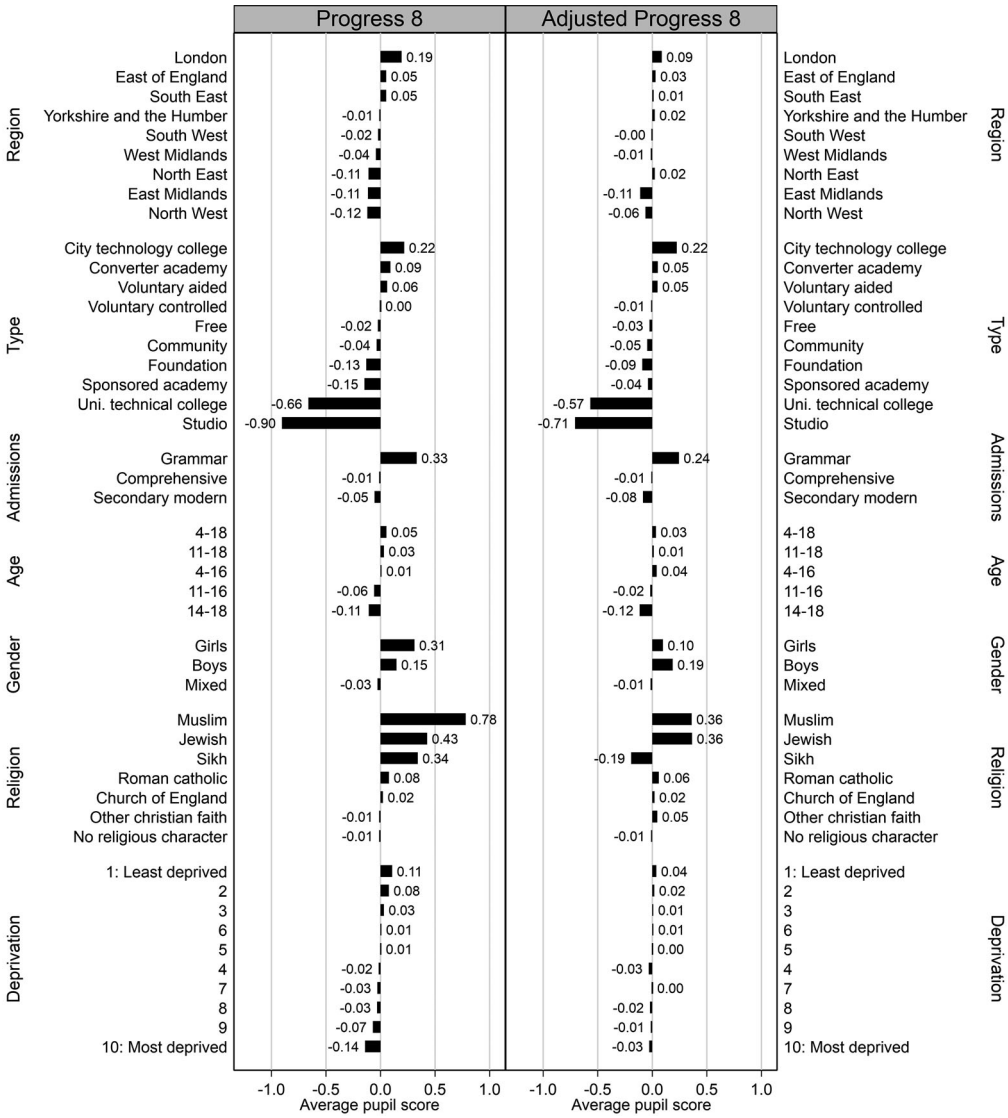

Figure 3. Average pupil Progress 8 and Adjusted Progress 8 scores by school characteristics
Note: The categories of each school characteristic are sorted by average pupil Progress 8 score.
There are only three city technology colleges.
There is only one Sikh school and eight Muslim schools.
The numbers of pupils and schools by school characteristic are given in Table S3.

progress pupil group. See Blanden *et al.* (2015) and Burgess (2014) for discussions of this 'London effect'. Now consider schools in the North East (152 schools; 5%), the region which shows almost the lowest average pupil progress according to Progress 8, with a score of −0.11. Under Adjusted Progress 8, this score increases to 0.02. Essentially, under Progress 8, schools in the North East are doubly disadvantaged not just by teaching relatively poor intakes, but also by disproportionately

teaching White British pupils. Both of these pupil characteristics are associated with below-average progress (Figure 1).

There are now a number of different school types in England (Hutchings & Francis, 2017; IPPR, 2017). Average pupil progress for many school types remains approximately the same when we move from Progress 8 to Adjusted Progress 8. However, for some school types, average pupil progress changes markedly. In particular, among converter academies (1,320 schools; 43% of all schools), average pupil progress drops from 0.09 to 0.05, while among sponsored academies (560; 18.1%), average pupil progress increases from −0.15 to −0.04. The superior performance of converter academies over sponsored academies is expected as only successful schools (as judged by Ofsted) are allowed to become converter academies while sponsored academies are usually set up to replace underperforming schools. Here the driving factor for the reduction in their apparent difference in performance is that converter academies teach a much lower percentage of poor pupils (20% eligible for FSM) than sponsored academies (40% eligible for FSM). Similarly, the very low average pupil progress seen in both university technical colleges (26 schools; 0.8%) and studio schools (30 schools; 1%) is substantially reduced once the types of pupils who tend to attend these schools is taken into account. Specifically, studio schools are disadvantaged by teaching a high percentage of SEN pupils (33%), while university technical colleges are disadvantaged by teaching a high percentage of boys (76%).

While nearly all schools in England are comprehensive (they do not in theory select on prior attainment), a small number of grammar schools (162 schools; 4.1%) use entrance examinations (House of Commons Education Committee, 2017). Schools in grammar school areas with no entrance examinations are referred to as secondary modern schools (117 schools; 3.5%). In terms of school admissions, according to Progress 8, pupils in grammar schools score, on average, a considerable 0.33 grades higher per subject than pupils nationally with the same prior attainment. However, under Adjusted Progress 8, the apparent benefit of attending a grammar school is reduced by almost a third: average pupil progress drops from 0.33 to 0.24. Grammar schools are especially advantaged by the low percentage of poor (6.8%) and to a lesser extent SEN pupils (5.6%) they teach, but are also advantaged by disproportionately teaching various high-progress ethnic groups. Interestingly, adjusting for pupil background leads secondary modern schools to appear less rather than more effective: average pupil progress drops from −0.05 to −0.08. The intuition for this result is that while secondary modern schools teach a much higher percentage of poor pupils than grammar schools (23.8% vs. 6.8%), they still teach lower percentages of poor pupils than schools nationally (26.6%). Adjusted Progress 8 takes this into account, leading to a slight lowering of average pupil progress.

Schools in England also vary somewhat in the age ranges they teach. Average pupil Progress 8 varies less dramatically by school age range and so we see only relatively small changes in average pupil progress when we move from Progress 8 to Adjusted Progress 8. According to both measures, there is some suggestion that pupils make more progress in schools teaching through to 18 than in schools teaching through to 16. However, more noticeable is the lower progress made by pupils in schools which teach from age 14 onwards. This last group is disproportionately university technical

colleges, studio schools and further education colleges, all of whose low progress was noted above.

While nearly all schools in England are mixed-sex, there are a small number of all-girls schools (209 schools; 6% of all schools) and all-boys schools (151 schools; 4% of all schools). Progress 8 suggests that pupils in single-sex schools, especially all-girls schools, make more progress than pupils in mixed-sex schools. However, average pupil progress in all-girls schools drops from 0.31 to 0.10 when we move from Progress 8 to Adjusted Progress 8. In contrast, the average pupil progress in all-boys schools increases from 0.15 to 0.19 and so the performance of all-boys schools now appears more impressive than that of all-girls schools. The reason for this change is that Adjusted Progress 8 adjusts for pupil gender whereas Progress 8 does not. Nationally, girls outperform boys (Figure 1). Thus, whereas Progress 8 compares girls in all-girls schools to girls and boys nationally, Adjusted Progress 8 only compares girls in all-girls schools to girls nationally. We note that single-sex schools are disproportionately grammar schools whose higher average pupil progress we have already reported.

A minority of schools in England follow a religious denomination (564 schools; 17.6%) (Long & Bolton, 2018). Progress 8 shows pupils in religious schools typically make more progress than those in schools with no religious character. Especially high progress is seen in the small number of Muslim (8 schools), Jewish (11 schools) and Sikh schools (1 school). However, the results for these schools change markedly when we turn to Adjusted Progress 8. In terms of Muslim schools, average pupil progress halves from 0.78 under Progress 8 to 0.36. The intuition for this drop is that these schools teach very high percentages of Indian (49.5%) and Pakistani (37%) pupils who also do not speak English as a first language (80.7%). These characteristics are nationally associated with making high progress (Figure 1). An even more extreme change is shown by the single Sikh school where average pupil progress changes from 0.34 under Progress 8 to −0.19 under Adjusted Progress 8. The large change seen here reflects that this school almost exclusively teaches Indian pupils (86%), one of the very highest progress ethnic groups. The average pupil progress for Jewish schools, in contrast, changes little. Here an analysis of the underlying data shows that accounting for ethnicity actually raises average pupil progress slightly, as Jewish pupils fall under the White British ethnic group which nationally underperforms. However, Jewish schools also teach relatively prosperous intakes and so the net effect is that their average pupil progress is nonetheless lowered when one also additionally accounts for FSM and deprivation.

Finally, the strong relationship between neighbourhood socioeconomic deprivation and pupil progress weakens substantially as we move from Progress 8 to Adjusted Progress 8. This result is not surprising as Adjusted Progress 8 adjusts for the deprivation of each pupil's neighbourhood, and in general most pupils in each school reside in neighbourhoods of similar deprivation to that of their school.

Discussion

In this article, we have explored whether school accountability systems should adjust for pupil demographic and socioeconomic background characteristics in their school

value-added models. We have critiqued the theoretical arguments for and against making these adjustments and examined their practical importance in the context of England's 'Progress 8' secondary school accountability system. Specifically, we modified Progress 8, which only adjusts for pupil prior attainment, to produce an 'Adjusted Progress 8' measure that additionally accounts for seven further pupil characteristics: age, gender, ethnicity, language, SEN, FSM and deprivation. We then compared Progress 8 and Adjusted Progress 8 in terms of schools' scores, ranks and classifications, and in terms of pupil average scores across a range of school characteristics.

The impact of adjusting Progress 8 for pupil background

Our results for Progress 8 show that adjusting for pupil background qualitatively changes many of the interpretations and conclusions one draws as to how schools in England are performing. For example, over a third of schools judged 'underperforming' according to the Progress 8 floor standard would no longer be judged underperforming according to Adjusted Progress 8. More generally, a fifth of schools would see their national league table positions change by over 500 places, which is substantial given there are only around 3,000 schools nationally. Pupil FSM and ethnicity prove the most important characteristics to consider. For example, the high average pupil progress seen in London more than halves when we adjust for pupil background and this is principally due to the high proportions of high-progress ethnic groups taught in London. In contrast, the low average pupil progress seen in the North East increases substantially after adjustment due to the disproportionately high proportions of poor pupils taught in this region. Other dramatic changes are seen for grammar schools and faith schools whose high average pupil progress reduces substantially once the educationally advantaged nature of their pupils is taken into account. In contrast, the low average pupil progress seen in sponsored academies increases once the disadvantaged nature of their pupils is recognised.

While our results quantify the average effectiveness of different school types, the data do not allow us to distinguish between different potential explanations as to why certain school types perform better than others. For example, the superior performance of grammar schools, even after adjusting for pupil background, may reflect genuinely higher-quality teaching in these schools, or it may reflect that it is easier to teach more able pupils if they are concentrated together, because lessons can be delivered at greater pace with less need to review and repeat.

Should we adjust Progress 8 for pupil background?

It seems clear from our results that the higher the proportion of disadvantaged pupils in a school, the more it will effectively be punished for the national underperformance of these pupil groups. It would therefore seem that value-added measures such as Progress 8, which ignore pupil background, give too much emphasis to schools rather than government or society as primarily responsible for these national differences in performance. In contrast, value-added measures such as Adjusted Progress 8, which account for pupil background, can be argued to stress that government and society

rather than schools are primarily responsible for these national differences. The decision to adjust can therefore be seen as a choice between two opposing views. However, there is no need to choose, especially as most would argue that schools, society, and government bear shared responsibility for the national differences that we see between different pupil groups. In the English context, it would seem that the government would therefore do better to publish and explain Progress 8 and an adjusted Progress 8 measure side-by-side to present a more informative picture of schools' performances.

Further methodological concerns with Progress 8

There are, however, other unusual methodological features to Progress 8 which raise further doubts as to its purported validity. In particular, Progress 8 follows a two-stage linear regression approach. However, the most commonly applied approach in the literature is to use multilevel models (Aitkin & Longford, 1986; Raudenbush & Willms, 1995; Goldstein, 1997, 2011; Teddlie & Reynolds, 2000; OECD, 2008; Reynolds *et al.*, 2014). We would argue that there are notable benefits of the multilevel approach to studying school effects. First, the approach is more robust to the confounding biases which will arise in the presence of any systematic sorting of more advantaged pupils into more effective schools (Castellano *et al.*, 2014). Second, the predicted school effects are so-called 'shrinkage' estimates which pull the estimated value-added scores of small schools towards the national average and therefore discourage unwarranted conclusions being drawn about the effectiveness of those schools where there is insufficient data to be statistical confident in making any such inferences (Goldstein, 2011). Third, the multilevel approach lends itself to the study of 'differential school effects', the notion that schools may make differential progress with different pupil groups (e.g. low prior attainers or particular ethnic groups) (Strand, 2016). Fourth, multilevel value-added models can easily be extended to incorporate separate scores on different academic subjects and across multiple cohorts of pupils (Leckie, 2018), facilitating richer summaries of school performance. The latter is particularly important given the instability of school effects over time and the considerable statistical noise surrounding estimates based on single cohorts of data (Leckie & Goldstein, 2011; Perry, 2016). Fifth, these models can also be adapted to account for the series of schools mobile pupils attend (Leckie, 2009), as opposed to the default approach of naïvely holding the final school attended accountable for the entirety of these pupils' education.

From 2006 to 2015, the government did use multilevel models to produce their school value-added measures (Leckie & Goldstein, 2017). The published school effects were 'shrinkage' estimates and in later years differential school effects were published for different academic subjects, cohorts and pupil groups. The decision to replace this with their current two-stage linear regression approach appears to be largely driven by a desire for simplicity; more complex approaches were argued difficult for practitioners to understand (DfE, 2010; Kelly & Downey, 2010b; Burgess & Thomson, 2013). However, there is no requirement for users to understand the technical details of the underlying statistical model in any given

approach, only how to interpret the resulting scores. Furthermore, similar and more complex approaches are being successfully applied in other schooling systems around the world (Leckie & Goldstein, 2017). What is therefore needed is a rigorous independent evaluation of the statistical strengths and weaknesses of the government's two-stage linear regression approach versus the multilevel modelling approach.

More general limitations of using school value-added measures for school accountability

Importantly, the methodological concerns we have expressed regarding Progress 8 are just a small subset of more general concerns with high-stakes testing and the use of school value-added models in school accountability systems, voiced both by academics (Foley & Goldstein, 2012; Amrein-Beardsley, 2014; Perry, 2016; Koretz, 2017; FFT Education Data Lab, 2018) and society more generally (NAHT, 2018; betterwithoutbaseline.org.uk; morethanascore.org.uk; vamboozled.com). Key concerns are that the tests fail to measure many important aspects of teaching (e.g. pupil engagement, curiosity and eagerness to learn), lead to a narrowing of the curriculum (e.g. they typically ignore arts, music, drama and other non-traditional academic subjects), result in teaching to the test, induce excessive pupil and teacher stress, create a culture of fear, tend to drive teachers out of the profession, lead to various gaming behaviours (e.g. excluding pupils from tests and cheating) and that the published scores are often presented with insufficient guidance, caveats or quantification of statistical uncertainty. Clearly these general concerns apply irrespective of whether pupil background is adjusted for or not, but the exact nature of, for example, the gaming behaviours employed by schools will likely change and experience shows us that it is often not possible to foresee the nuances of such changes in advance (Foley & Goldstein, 2012). Perhaps, most worryingly, there is still very little research demonstrating the actual improvement to pupil learning that school accountability via pupil test scores and school value-added measures is meant to bring about (NFER, 2018). More research is needed in this area, but the uniform national implementation of reforms to the accountability system makes such evaluations challenging in practice (Burgess *et al.*, 2013; Goldstein & Leckie, 2016).

Our own view is that the results presented here, coupled with these more general concerns, raise serious doubts about not just Progress 8 but test-based school accountability more generally. In terms of Progress 8, the types of automated data-driven decision-making that the government currently aspires to, whereby schools falling below a single floor standard are declared underperforming, cannot be supported by the data. Our view is that, for school accountability purposes, the most school value-added measures can be used for is as 'screening devices' to choose schools for careful sensitive further investigation (Foley & Goldstein, 2012). However, we believe that a better use is simply as tools for school self-evaluation, where they can potentially help inform schools on the policies and practices which help different pupil groups to reach their potential, but further discussion of this is outside the scope of the present article.

Acknowledgements

We are grateful for helpful comments and suggestions from Dave Thomson at Fischer Family Trust and from the Editors and the anonymous reviewers. This research was funded by UK Economic and Social Research Council grant ES/R010285/1.

References

Aitkin, M. & Longford, N. (1986) Statistical modelling issues in school effectiveness studies, *Journal of the Royal Statistical Society, Series A*, 149(1), 1–43.

Amrein-Beardsley, A. (2014) *Rethinking value-added models in education: Critical perspectives on tests and assessment-based accountability* (Abingdon, Routledge).

BBC (2018, May 25) League tables changes 'toxic' for poor white schools.

Blanden, J., Greaves, E., Gregg, P., Macmillan, L. & Sibieta, L. (2015) *Understanding the improved performance of disadvantaged pupils in London*. Social Policy in a Cold Climate, Working Paper No. 21. (London, Centre for Analysis of Social Exclusion, London School of Economics).

Burgess, S. (2014) *Understanding the success of London's schools*. Bristol Centre for Market and Public Organisation, Working Paper.

Burgess, S. & Thomson, D. (2013) *Key stage 4 accountability: Progress measure and intervention trigger*. Bristol Centre for Market and Public Organisation, Report.

Burgess, S., Wilson, D. & Worth, J. (2013) A natural experiment in school accountability: The impact of school performance information on pupil progress, *Journal of Public Economics*, 106, 57–67.

Castellano, K. E. & Ho, A. D. (2013) *A practitioner's guide to growth models* (Washington, D.C., Council of Chief State School Officers).

Castellano, K. E., Rabe-Hesketh, S. & Skrondal, A. (2014) Composition, context, and endogeneity in school and teacher comparisons, *Journal of Educational and Behavioral Statistics*, 39, 333–367.

Crawford, C., Dearden, L. & Meghir, C. (2007) *When you are born matters: The impact of date of birth on child cognitive outcomes in England* (London, Centre for the Economics of Education, London School of Economics and Political Science).

DfE (2010) *The importance of teaching: The Schools White Paper 2010* (London, Department for Education).

DfE (2018a) *Permanent and fixed period exclusions in England: 2016 to 2017* (London, Department for Education).

DfE (2018b) *Progress scores for key stage 4: School and college performance tables* (London, Department for Education).

DfE (2018c) *Secondary accountability measures: Guide for maintained secondary schools, academies and free schools* (London, Department for Education).

EPI (2017) *The introduction of Progress 8* (London, Education Policy Institute).

FFT Education Data Lab (2018) *Value added measures in performance tables: A recap of the main issues for secondary schools* (London, FFT Education Data Lab).

Foley, B. & Goldstein, H. (2012) *Measuring success: League tables in the public sector* (London, British Academy).

Goldstein, H. (1997) Methods in school effectiveness research, *School Effectiveness and School Improvement*, 8, 369–395.

Goldstein, H. (2011) *Multilevel statistical models* (4th edn) (Chichester, Wiley).

Goldstein, H. & Leckie, G. (2016) Trends in examination performance and exposure to standardised tests in England and Wales, *British Educational Research Journal*, 42, 367–375.

House of Commons Education Committee (2017) *Evidence check: Grammar schools* (London, House of Commons Education Committee).

Hutchings, M. & Francis, B. (2017) *Chain effects 2017: The impact of academy chains on low-income students* (London, The Sutton Trust).

Ilie, S., Sutherland, A. & Vignoles, A. (2017) Revisiting free school meal eligibility as a proxy for pupil socio-economic deprivation, *British Educational Research Journal*, 43, 253–274.

IPPR (2017) *Tech Transitions. UTCs, studio schools, and technical and vocational education in England's schools* (London, Institute for Public Policy Research).

Kelly, A. & Downey, C. (2010a) *Using effectiveness data for school improvement: Developing and utilising metrics* (Abingdon, Routledge).

Kelly, A. & Downey, C. (2010b) Value-added measures for schools in England: Looking inside the 'black box' of complex metrics, *Educational Assessment, Evaluation and Accountability*, 22, 181–198.

Koretz, D. (2017) *The testing charade: Pretending to make schools better* (Chicago, IL, University of Chicago Press).

Leckie, G. (2009) The complexity of school and neighbourhood effects and movements of pupils on school differences in models of educational achievement, *Journal of the Royal Statistical Society, Series A*, 172, 537–554.

Leckie, G. (2018) Avoiding bias when estimating the consistency and stability of value-added school effects using multilevel models, *Journal of Educational and Behavioral Statistics*, 43, 440–468.

Leckie, G. & Goldstein, H. (2009) The limitations of using school league tables to inform school choice, *Journal of the Royal Statistical Society, Series A*, 172, 835–851.

Leckie, G. & Goldstein, H. (2011) Understanding uncertainty in school league tables, *Fiscal Studies*, 32, 207–224.

Leckie, G. & Goldstein, H. (2017) The evolution of school league tables in England 1992–2016: 'Contextual value-added', 'expected progress' and 'progress 8', *British Educational Research Journal*, 43, 193–212.

Long, R. & Bolton, P. (2018) *Faith schools in England: FAQs* (London, House of Commons Library).

NAHT (2018) *Improving school accountability* (London, National Association of Head Teachers).

NFER (2018) *What impact does accountability have on curriculum, standards and engagement in education? A literature review* (National Foundation for Educational Research, Slough).

OECD (2008) *Measuring improvements in learning outcomes: Best practices to assess the value-added of schools* (Paris, Organisation for Economic Co-operation and Development Publishing & Centre for Educational Research and Innovation).

Perry, T. (2016) English value-added measures: Examining the limitations of school performance measurement, *British Educational Research Journal*, 42, 1056–1080.

Perry, T. (2018) 'Phantom' compositional effects in English school value-added measures: The consequences of random baseline measurement error, *Research Papers in Education*, 34, 239–262.

Raudenbush, S. W. & Willms, J. (1995) The estimation of school effects, *Journal of Educational and Behavioral Statistics*, 20, 307–335.

Reynolds, D., Sammons, P., De Fraine, B., Van Damme, J., Townsend, T., Teddlie, C. *et al.* (2014) Educational effectiveness research (EER): A state-of-the-art review, *School Effectiveness and School Improvement*, 25, 197–230.

Sammons, P. (1995) Gender, ethnic and socio-economic differences in attainment and progress: A longitudinal analysis of student achievement over 9 years, *British Educational Research Journal*, 21, 465–485.

Selfridge, R. (2018) *Databusting for schools: How to use and interpret education data* (London, Sage).

Strand, S. (2014) Ethnicity, gender, social class and achievement gaps at age 16: Intersectionality and 'getting it' for the white working class, *Research Papers in Education*, 29, 131–171.

Strand, S. (2016) Do some schools narrow the gap? Differential school effectiveness revisited, *Review of Education*, 4, 107–144.

Strand, S., Malmberg, L. & Hall, J. (2015) *English as an additional language (EAL) and educational achievement in England: An analysis of the National Pupil Database* (London, Educational Endowment Fund).

Teddlie, C. & Reynolds, D. (2000) *The international handbook of school effectiveness research* (London, Psychology Press).
TES (2018, April 13) Exclusive: Progress 8 'penalises schools in white working class communities', study shows.
Timmermans, A. C. & Thomas, S. M. (2015) The impact of student composition on schools' value-added performance: A comparison of seven empirical studies, *School Effectiveness and School Improvement*, 26, 487–498.
West, A. (2010) High stakes testing, accountability, incentives and consequences in English schools, *Policy & Politics*, 38, 23–39.
Wilson, D., Burgess, S. & Briggs, A. (2011) The dynamics of school attainment of England's ethnic minorities, *Journal of Population Economics*, 24, 681–700.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Table S1. Pupil- and school-level summary statistics for key stage 2, Attainment 8, Progress 8 and Adjusted Progress 8.
Table S2. Distribution of pupils by pupil demographic and socioeconomic characteristics.
Table S3. Distribution of pupils and schools by school characteristics.
Table S4. Model results for Progress 8 and Adjusted Progress 8 linear regression models.
Figure S1. Distribution of pupil Attainment 8 scores and pupil KS2 scores.
Figure S2. Scatterplot of pupil Attainment 8 scores against pupil KS2 scores.
Figure S3. Distribution of pupil Progress 8 and school Progress 8 scores.
Figure S4. Average pupil KS2 and Attainment 8 scores by pupil characteristics.
Figure S5. Average pupil KS2 and Attainment 8 scores by school characteristics.